Content Extraction Enhancements For Target Analytics:

SMS Text Messages: A Goldmine to Exploit

9 June, 2011

Presenters: [redacted]

[redacted]

With

[redacted]

Work funded by T1221 Center for Content Extraction

Performed in Collaboration with

[redacted] T1221 Center for Content Extraction

[redacted] T132 Dishfire

(U) OUTLINE

- (U) Introduction & Some Statistics
- (U) Missed Call Messages
- (U) MilkBone QFD “Demo”
- (U) Where Next?

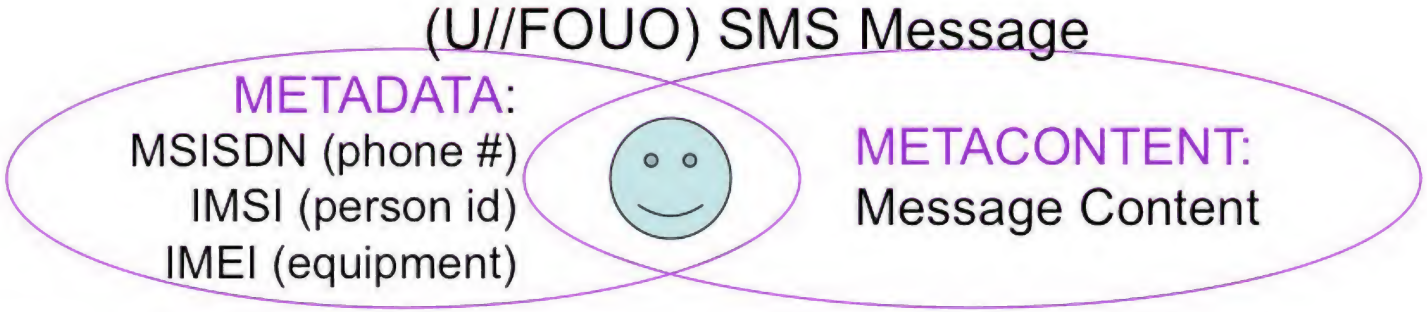
(U)SMS (Short Message Service) some stats

- (U) (May 2011): Mobile phone subscriptions have reached 5.3 billion, 77% of the world population. Growth led by China and India.
- (U) 500 million people accessed mobile internet worldwide in 2009. Usage is expected to double in 5 years. 1/2011: 200 million users access Facebook using mobile.
- (U)(Oct. 2010) Many mobile Web users are mobile-only (rarely use desktop, laptop or tablet to access the Web|). Mobile-only in Egypt is 70%, India 59% and US 25%. Mobile penetration in the developing world is now at 68%,
- (U) SMS is still king of mobile messaging – 6.1 trillion messages sent in 2010 (200,000 text messages per second) and is expected to exceed 10 trillion in 2013 (1.8 trillion sent in 2007). Most number of texts are sent in the Philippines and US.
- (U) Mobile phone providers in developing countries increasingly use the mobile phone for health services and banking (International Telecommunications Union)
- (U) Many mobile web users do not have a bank account (India 57%). Gartner predicts that the number 1 service in 2010 will be money transfer using SMS. Estimate 2009 55 million users and various organizations predict doubling every year estimate 2013 around 5 million user). Initiatives to bank the unbanked.
- (U) The typical mobile subscriber sends and receives more SMS text messages than telephone calls. The average U.S. mobile customer sent or received 357 text messages in 2008 (a 450% increase over 2006) and placed/received 204 calls. In 2010, the average American teen sent or received 3,339 texts per month, > 6 per hour.
- (U) 2008 estimate of text message usage among wireless subscribers: Russia – 88%, UK – 76%, China – 72%, Brazil – 60%, USA – 53%

(U) SMS Message Components

- (U) SMS Metadata
  - (U) IMSI: International Mobile Subscriber Identity (most frequent)
  - (U) MSISDN: Mobile Subscriber Integrated Services Digital Network Number, i.e., phone number
  - (U) IMEI: International Mobile Equipment Identity
  - (U) SME: Short Message Entity (entities which can send & receive messages)
- Content
  - (U) Typed Text Message
    - (U) User entered
    - (U) System Generated
      - (U) Useful (personal) [Ham]
      - (U) Spam

(U) Why?



- (S//REL) Metadata + Content of System Generated Text Messages leads to analytic gems => content derived metadata
- (S//SI//REL) Such gems often are not in current metadata stores and would enhance current analytics: contact chaining, geolocation, alternative identifiers (including DNI & DNR links), travel, finance
- (S//REL) SMS: Rich data set, high impact. Usage is increasing. Features & Notifications available on mobile phones are increasing → rich data set awaiting exploitation.

(U) PREFER

- (U//FOUO) Identifies types of automated messages
- (U//FOUO) Extracts entities from SMS content daily:
- (S//REL) Results presented averaged over 30 days (April 2011)
  - 194,184,810 - sms messages per day (not deduped)
  - 184,794,279 - DISHFIRE message tags
  - 188,299,963 - PREFER text slice decoded
- (S//REL) PREFER operational on DISHFIRE servers since January 2008, inserting content derived tags into xml output. First major utilization, SPYDER 2008 for selected content.

(U) How Does PREFER Fit

Collection

Discover the Targets

DISHFIRE
(SMS repository)

Extract Useful Information

PREFER Extraction:
-message recombination
-"metadata" from content

Follow the Targets

Existing Tools/Databases

MetaContent QFDs

DISHFIRE
SPYDER
SEDB
FASTCOPE Innovations

MILKBONE

(U//FOUO) PREFER

Identification & Extraction April 2011

(S//SI//REL) 194 Million Messages Collected by DISHFIRE per Day, Including

- (S//SI//REL) VCARDS → names+; (113,672 average extracted daily)
  sometimes DNI link (email) to DNR (telephony) as well as images
- (S//SI//REL) Geocoordinates (76,142 daily avg; hex-encoded 10,432)
  - Requests by people for route info
  - Setting up meetings at a location
  - Tracking information: e.g., [redacted] (12,809)
  - Comma Separated Formats (33,020)
- (S//SI//REL) Missed Calls → contact chaining (5,058,114)
- (S//SI//REL) SIM Card Changes → IMSI/IMEI links (6,017,901)
- (S//SI//REL) Roaming information → border crossings (1,658,025)
- (S//SI//REL) Travel (5,314)
  - Itinerary including multiple flights
  - Changes: cancellations, reschedules, delays
- (S//SI//REL) Financial Transactions:
  - Credit card transactions: correlate credit cards to individuals (61,488)
  - Money transfers (social networks) – Phone to Phone (630,846)
  - Track financial information (account activity – bank transaction) (115,480)
- (S//SI//REL) Passwords (pending); Other Requests?

[redacted]